Phylogenetic Analysis Using Bayesian Model

Wei Lu, Ph.D. ,P.Eng. Member, IEEE
Department of Computer Science
Keene State College. USNH
Keene, NH USA
e-mail: wlu@keene.edu

Mike Hanrahan, Ph.D.
Department of Computer Science
Keene State College, USNH
Keene, NH USA
e-mail: mhanraha@keene.edu

***Abstract* - Bayesian inference has been widely applied for phylogenetic and phyloinformatic analysis in recent years. In this paper we build up a high performance computing platform using OpenMPI with free-cost open source Ubuntu Linux operating systems, and then apply the Bayesian inference model to construct a phylogenetic tree of various biological species based on similarities and differences in their physical and genetic characteristics. A case study is also conducted and the experimental evaluation results to show that our cluster platform has achieved a good performance in the terms of time complexity when analyzing the molecular data using Bayesian model, leading a pilot inter-disciplinary Bioinformatics education program in between Computer Science and Biology.**

Keywords - linux, phylogenetics, phyloinformatics, operating system

I. INTRODUCTION

Phylogenetic analysis is currently playing a significant role in nearly all fields of bioinformatics. With more and more data obtained by biological researchers, the size of phylogenetic analyses and the scope of inferences have increased dramatically. Researchers typically employ a greater set of tools for phylogenetic analyses compared to previous years in which one or two simple programs were applied. Although the biologist receives a large benefit from these diverse sets of computing programs, the resource requirements become significant. Not only do they require large areas for the large data sets required, but increased time and resources must be allowed for software installation, configuration, and maintenance.

Over the last decades, supercomputing techniques and more recently cloud computing has been widely applied to solve the time consuming tasks in bioinformatics. However to participate in these resources, the institutions must continually apply for grant to reserve their shared research space. And if received, the user space and computer time is many times reserved only for the research faculty involved. In many institutions this prohibits especially undergraduate students from participating in the research. Now, however, the availability of cheap multi-core processors makes it possible for many even-moderately sized labs to build routinely small clusters including a group of phylogenetic workstations. This is important for phyloinformatic research because it depends largely on processing power by which highly parallel analyses are spread across many computers and processors. As the phyloinformatic pipelines become more and more sophisticated, a careful standardization of operating systems across multiple smaller networked clustered computers allows even the small lab to economically create a user friendly computing environment in which complex and time-consuming analyses, and compiling, installing, and configuring software can be conducted easily without having to rely on off-site expensive super-computers and cloud storage.

In this research, by collaborating with our colleagues in the filed of Biology we address this issue and build a lightweight Ubuntu Linux-based cluster which will be used for phylogenetic and phyloinformatic research, and then conduct an experimental study to evaluate the effectiveness of the cluster by analyzing a set of molecular data using the Bayesian model, a very popular tool in the field of phyloinformatics. It should be noted that existing Linux distributions such as Bio-Linux [1] and SciBuntu [2] represent some typical examples for building a cluster platform; however these distributions are aimed at a more general areas of bioinformatics and thus do not include many of the software packages that are now standard tools for phylogenetic analysis. One of the major goals of our work is to prompt the engineering education program in the traditional small-size liberal arts school and the research on the Bioinformatics represent a good pilot study for this inter-disciplinary research including the faculty members of both Computer Science and Biology. The main contributions of this paper include: (1) building and setting up a high performance networked computing cluster of multi-core computers using Ubuntu Linux which has a very low implementation cost compared to other existing commercial HPC (High Performance Computing) platforms; (2) applying Bayesian inference model on the cluster and evaluating its effectiveness to analyze Phylogenetic data;

The rest of the paper is structured as follows: Section II presents the framework and our implementation of an Ubuntu Linux based cluster. Section III introduces the basic idea of Bayesian inference model and Markov Chain Monte Carlo (MCMC) techniques applied for phylogenetic analyses. Section IV outlines the experimental evaluation settings and the results of a comparative study on analyzing molecular

978-1-4799-5233-5/14/$31.00 ©2014 IEEE

data. Finally, Section V provides concluding remarks and an outline of intended future research.

II. Implementation of the Ubunbu Linux Cluster

The theory of HPC is to divide a complex problem into smaller component tasks that can be worked on simultaneously, thus solving the problem more quickly. A typical HPC computing system usually consists of one master node and multiple child computing nodes connected via standard network interconnects such as a gigabit switch. In this implementation, we chose to run all of the nodes on open-source Ubuntu Linux as it provided substantial savings over proprietary operating systems as well as a low cost solution for the cluster deployment and implementation.

The master node of the cluster acts as a server for the Network File System (NFS), job-scheduling, and security. It also provides a gateway to end-users. The master node assigns each of the child computing nodes with one or more tasks to perform. The larger task is split into sub-functions. As a gateway, the master node allows the users to gain access to the compute nodes. The sole task of the child computing nodes is to execute assigned tasks in parallel. Access to client nodes is provided via remote connections through the master node.

Figure 1 illustrates the architecture of our cluster which we began with five nodes but which can be easily expanded as computing needs increase. The cluster consists of 5 Dell Optiplex 790 desktops with:

(1) 2nd Gen Intel Core i3-2120 Processor (Dual Core, 3.30GHz)
(2) 4GBDual Channel DDR3 SDRAM at 1333MHz
(3) 500GB 7,200 RPM 3.5" SATA, 6.0Gb/s Hard Drive with 8MB Cache

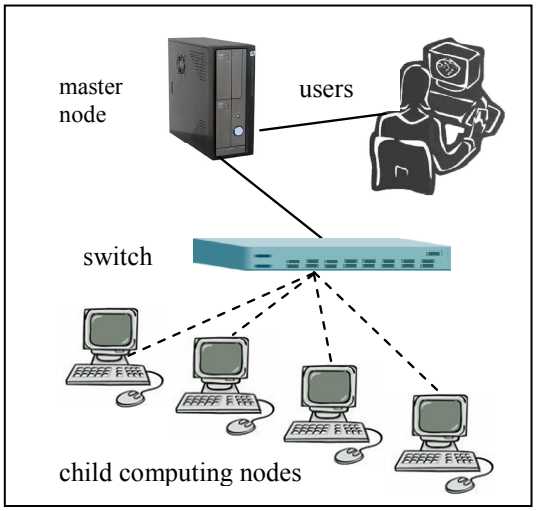

Figure 1. Architecture of Ubuntu Linux Cluster

The nodes are connected via gigabit Ethernet connectors to a 5-port Netgear GS-105 5-port Giga Switch.

The cluster is built based on an OpenMPI technology [3] in which the master compute node has two network adapters: one is connected to the internal network and the other one is connected to external network. The operating system installed in the 5 computers is 64-bit Ubuntu Linux Server (12.04 LTS edition). Please see [13] online for the details on configuring and setting up the HPC platform based on Ubuntu Linux.

III. Bayesian Inference Model

It is well documented that in evolutionary biology organisms sharing a common origin mutate over generations. In recent years, Phylogenies has been able to reconstruct the evolutionary history of these mutations. This is possible because of the better access to DNA sequences available now by utilizing the resources and massive computing power of super computers. A typical example is one which uses Bayesian inference techniques [5] [6] [7]. These techniques have been widely applied in the filed of phylogeny due to their ability to incorporate prior information and their ability to the more easily interpreted results using advanced computing machinery when compared to other inference approaches.

In phylogeny, Bayesian inference is typically applied to create the posterior distribution for a parameter based on the prior probability of that parameter and the likelihood of the DNS sequence data. In the basic Bayesian theory we have:

$$p(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{p(D)}$$

in which $p(\theta \mid D)$ is the posterior probability distribution of phylogenetic trees; $p(\theta)$ is the prior probability distribution; $p(D)$ is the marginal probability of the data in which we have:

$$p(D) = \int_{\Theta} p(D \mid \theta) p(\theta) d\theta$$

and the parameter space $\Omega = (\Psi, \Phi)$ includes the set of all possible phylogenetic trees which is denoted by Ψ and the set of all likelihood model parameters denoted by Φ.

An instance of phylogenetic trees can be represented as follows:

$$\psi = (\tau, \beta)$$

in which τ is the topology of the tree and β is branch length of the tree. Given the model $p(D \mid \theta)$, D and φ, we then have:

$$p(\tau, \beta, \varphi \mid D) = \frac{p(D \mid \tau, \beta, \varphi) p(\tau, \beta, \varphi)}{\sum_{\tau} \int_{B} \int_{\Phi} p(D \mid \tau, \beta, \varphi) p(\tau, \beta, \varphi) d_{\varphi} d_{\beta}}$$

Bayesian inference depends heavily on the posterior probability and in most cases, the summation and integrals are difficult to be calculated analytically. Therefore, the Markov chain Monte Carlo (MCMC) algorithm is proposed

to approximate the posterior probabilities of trees, which is illustrated as following:

Step 1: randomly choose the initial parameter

$$\theta^{(0)} = (\psi^{(0)}, \varphi^{(0)})$$

Step 2: set a loop and set the current state

$$\theta^{(i)} = (\psi^{(i)}, \varphi^{(i)})$$

(1) make $\psi^{(i)}$ unchanged , based on the Markov chain q_1 a new parameter $\varphi^{(*)}$ is generated.

(2) According to $R = \min(1, \frac{p(D \mid \theta^*) p(\theta^*)}{p(D \mid \theta) p(\theta)})$ and the random number generator, the algorithm makes a decision if a new state will be accepted or not: *if R = 1, new state will be accepted; if R<1 a random number a will be generated and if R > a, the new state will be accepted, otherwise the new state will be rejected*.

(3) Once a new state is accepted by MCMC, we have $\varphi^{(i+1)} = \varphi^{(*)}$, otherwise, we have $\varphi^{(i+1)} = \varphi^{(i)}$

(4) make $\varphi^{(i+1)}$ unchanged , based on the Markov chain q_2 a new tree model $\psi^{(*)}$ is generated.

(5) According to $R = \min(1, \frac{p(D \mid \theta^*) p(\theta^*)}{p(D \mid \theta) p(\theta)})$ and the random number generator, the algorithm makes a decision if a new state will be accepted or not: *if R = 1, new state will be accepted; if R<1 a random number a will be generated and if R > a, the new state will be accepted, otherwise the new state will be rejected*.

(6) Repeat (1) to (5) *n* times until the loop terminated and we then analyze the final result.

IV. EXPERIMENTAL EVALAUTION

In the experimental evaluation, we test and evaluate our cluster platform using the Bayesian inference model with a DNA sequence dataset including about 140 species in order to select a wide range of phylogenetic and evolutionary models. The Markov chain Monte Carlo (MCMC) algorithm has been implemented to estimate the posterior distribution of model parameters [10] [11].

To execute the evaluation and conduct a simple Bayesian MCMC analysis of phylogeny, we read the *.nex* data file into the program and then setup the evolutionary model using a set of parameters specified below. We then run the analysis and summarize the samples.

The specific steps of the evaluation are illustrated as follows [12]:

(1) The *infile.nex* data is brought into the program.
(2) Using the Bayesian reference mode, the GTR model with gamma-distributed rate variation is selected.
(3) The number of generations is set to 5000000 and the sampling frequency to 1000. This will cause the program to execute until the total number of generations reaches.
(4) Summarize the parameter values and generate a table with summaries including the mean, mode, and the 95% credibility interval of each parameter.
(5) Summarize the trees and then output a cladogram with the posterior probabilities for each split and a phylogram with mean branch lengths.

The experimental evaluation shows that it takes 78781.43 seconds of CPU time to complete total number of 5,000,000 generations. An example of the phylogram of evolutionary tree is illustrated as follows in Figure 2.

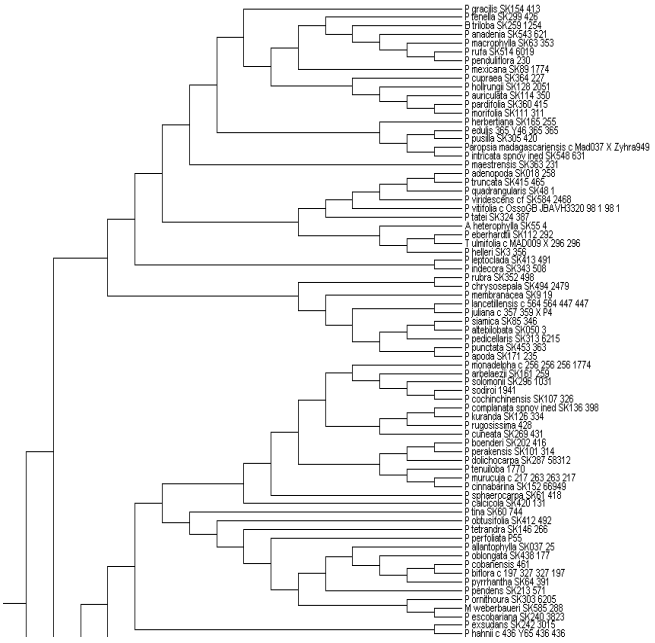

Figure 2. An Example of Evolutionary Tree

V. CONCLUSIONS

In this paper we construct an economical, lightweight, high performance computing cluster platform using the OpenMPI technologies with Ubuntu Linux operating systems for phylogenetics and phyloinformatics. The Bayesian inference technologies are applied on the cluster. A dataset including the real DNA sequence for around 140 species is applied to evaluate the effectiveness of the cluster and experimental results show that an evolutionary tree has been constructed successfully after 5,000,000 generations. In the future we will propose new models for phylogenetic analysis based on the current cluster platform and will implement new visualization software for constructing the generational trees.

ACKNOWLEDGMENT

We would like to acknowledge and especially appreciate Dr. Kristen Porter-Utley, Associate Professor of Biology, Keene State College, for her contributions to the DNA sequence data and for providing the corresponding parameters for the evolutionary model which we used in the experimental evaluation of the cluster platform.

REFERENCES

[1] D. Field, B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston, 2006. Open Software for biologists: from famine to feast. Nature Biotechnology 24, 801 - 803.

[2] U. Anjar Scibuntu: Ubuntu Linux for scientists. Ver. 0.4-beta. 2013. Available from http://scibuntu.sourceforge.net/index.html

[3] OpenMPI: Open Source High Performance Computing. Available from http://www.open-mpi.org/

[4] B. Mau, "Bayesian phylogenetic inference via Markov chain Monte carlo methods", PhD Dissertation, University of Wisconsin, Madison, 1996.

[5] B. Mau, M. Newton and B. Larget, Bayesian phylogenetic inference via Markov chain Monte carlo methods. Biometrics, vol. 55, pp. 1–12, 1999.

[6] Z. Yang, and B. Rannala, Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte carlo method. Mol. Biol. Evol., 1vol. 4, pp. 717–724, 1997.

[7] C.J. Geyer, Markov chain Monte Carlo maximum likelihood. In Keramidas (ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.. Interface Foundation, Fairfax Station, pp. 156–163, 1991.

[8] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82, pp. 711–732, 1995.

[9] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, pp. 97–109, 1970.

[10] J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics. 2001 Aug;17(8):754-5.

[11] MrBayes: Bayesian Inference of Phylogeny Available from http://mrbayes.sourceforge.net/

[12] S. Krosnick,, K. Porter-Utley, J.M. MacDougal, P. Jørgensen, and L. McDade. "New Insights into the Evolution of the Tiny Passionflowers: Phylogenetic Relationships in Passiflora subgenus Decaloba", accepted by Systematic Botany, 2013.

[13] Keene State HPC cluster, available in Mar. 2014 from http://bio.keene.edu/ASEE2014-Appendix.pdf